Information theory can help quantify the potential of new phenotypes to originate as exaptations

Wagner, Andreas

Abstract: Exaptations are adaptive traits that do not originate <jats:italic>de novo</jats:italic> but from other adaptive traits. They include complex macroscopic traits, such as the middle ear bones of mammals, which originated from reptile jaw bones, but also molecular traits, such as new binding sites of transcriptional regulators. What determines whether a trait originates <jats:italic>de novo</jats:italic> or as an exaptation is unknown. I here use simple information theoretic concepts to quantify a molecular phenotype's potential to give rise to new phenotypes. These quantities rely on the amount of genetic information needed to encode a phenotype. I use these quantities to estimate the propensity of new transcription factor binding phenotypes to emerge <jats:italic>de novo</jats:italic> or exaptively, and do so for 187 mouse transcription factors. I also use them to quantify whether an organism's viability in one of 10 different chemical environment is likely to arise exaptively. I show that informationally expensive traits are more likely to originate exaptively. Exaptive evolution is only sometimes favored for new transcription factor binding, but it is always favored for the informationally complex metabolic phenotypes I consider. As our ability to genotype evolving populations increases, so will our ability to understand how phenotypes of ever-increasing informational complexity originate in evolution.

DOI: https://doi.org/10.3389/fevo.2020.564071

Information theory can help quantify the potential of new phenotypes to originate as exaptations

Andreas Wagner[1,2,3*]

[1]University of Zurich, Institute of Evolutionary Biology and Environmental Studies, Zurich, Switzerland,

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland,

[3]Santa Fe Institute, Santa Fe, New Mexico, USA

*** Correspondence:**
Corresponding Author
andreas.wagner@ieu.uzh.ch

Keywords: innovation, novelty, evolution, co-option, metabolism, transcription, exaptation

6000 words, 3 figures

Abstract

Exaptations are adaptive traits that do not originate de novo but from other adaptive traits. They include complex macroscopic traits, such as the middle ear bones of mammals, which originated from reptile jaw bones, but also molecular traits, such as new binding sites of transcriptional regulators. What determines whether a trait originates *de novo* or as an exaptation is unknown. I here use simple information theoretic concepts to quantify a molecular phenotype's potential to give rise to new phenotypes. These quantities rely on the amount of genetic information needed to encode a phenotype. I use these quantities to estimate the propensity of new transcription factor binding phenotypes to emerge *de novo* or exaptively, and do so for 187 mouse transcription factors. I also use them to quantify whether an organism's viability in one of 10 different chemical environment is likely to arise exaptively. I show that informationally expensive traits are more likely to originate exaptively. Exaptive evolution is only sometimes favored for new transcription factor binding, but it is always favored for the informationally complex metabolic phenotypes I consider. As our ability to genotype evolving populations increases, so will our ability to understand how phenotypes of ever-increasing informational complexity originate in evolution.

Introduction

Evolution creates many traits from previously existing traits rather than *de novo*. This notion is as old as Charles Darwin's *Origin of Species*, where Darwin pointed to examples that include the lungs of vertebrates, which are homologous to the swim bladders of fish, and respiratory organs of some barnacles, which he thought evolved from egg-retaining structures he called ovigerous frenae (Darwin, 1872, p 176-177). Multiple morphological examples like these were already known in the 19th century, and during the 20th century, numerous molecular examples were also discovered. They include the crystallins, proteins that allow eye lenses to refract light. Many of them are co-opted from proteins that include metabolic enzymes and heat shock proteins (Piatigorsky and Wistow, 1989; Piatigorsky, 1998). Other examples include the regulatory circuits formed by Hox genes, which pattern the main body axis in many organisms and have been co-opted for multiple other purposes, among them vertebrate limb development (True and Carroll, 2002). Since the late 20th century such co-opted traits have been called exaptations (Gould and Vrba, 1982).

Not all exaptations are created equal. Some exaptations require substantial change of an ancestral trait, like the change that transform a forelimb into a bat's wing. Others, like crystallins, require as little change as expressing a gene in a new location like the eye lens. Yet others may require no genetic change at all. Examples include enzymes that catalyze not just one main reaction but multiple side reactions – such as antibiotic resistance proteins that cleave one 'native' antibiotic and multiple others (Khersonsky and Tawfik, 2010). Such latent beneficial traits only require the right (antibiotic containing) environment to become adaptive, illustrating that environmental change can play an important role in the origin of exaptations.

Many co-opted traits are bifunctional, serving their old and a new role simultaneously. For example, the fore- and hind-limbs of flying squirrels (Pteromyini) have been altered to help support the patagium, the membrane that allows these animals to glide between trees. At the same time, these limbs still allow walking and running (Thorington and Santana, 2007). Some such bifunctional intermediates undergo further refinement or secondary adaptation (Gould and Vrba, 1982). Take feathers, which probably originated to keep a body warm (Norell and Xu, 2005). Their shape and structure needed to be transformed before birds could fly, such that modern pennaceous feathers serve their new role in flight better than their original role in thermal insulation. In yet other traits, secondary adaptation causes a complete loss of the original role. A molecular example is the lactalbumin protein, which helps mammals synthesize lactose. It derives from lysozyme, an ancient enzyme, but has completely lost its original bacteriocidal function (Qasba and Kumar, 1997). Another example involves the middle ear bones that transduce sound in mammals but are derived from reptile jaw bones. One of them is the incus,

which is derived from the quadrate bone of the reptilian upper jaw. In therapsids – extinct reptiles that include the ancestors of today's mammals – the quadrate functioned both as part of the jaw joint and as part of the auditory system, but has since lost its original function in the jaw joint (Fay et al., 2004; Luo, 2007). Most exaptations will have such bifunctional intermediates, because the simultaneous loss of an old phenotype and gain of a new phenotype is more difficult and rare than a gradual shift between phenotypes. I will thus focus on the transition between a trait and a bifunctional intermediate.

The importance of exaptations in biological evolution has been defended largely on the basis of known examples. However, it would be useful to *quantify* how likely any one trait is to originate *de novo* or as an exaptation from another trait. And it would be useful to compare traits in this regard. Is it harder to evolve the Panda's thumb from a sesamoid bone, or our inner ear bones from a reptilian jaw? Questions like this cannot currently be answered for morphological traits, because their genetic basis is too complex, involving hundreds of genes with poorly understood interactions. However, the answer may be within reach for simpler, molecular phenotypes whose genetic basis is better understood, and that can originate even on the short time scales of laboratory evolution. Examples include a cell's ability to thrive in novel chemical environments, or to regulate genes in new ways (Blount et al., 2008; Dhar et al., 2011; Dhar et al., 2013; de Visser and Krug, 2014; Palmer et al., 2015; Toll-Riera et al., 2016).

Building on previous work (Wagner, 2017), I here suggest a method to quantify a phenotype's likelihood to evolve *de novo*. I apply the method to two simple kinds of molecular phenotype. The first of them is the ability of transcription factors to bind short DNA sequences, which is essential for gene regulation. The evolution of new transcription factor binding sites helps bring forth novel traits as different as new color phenotypes (Gompel et al., 2005) and novel body structures (Prud'homme et al., 2007; Guerreiro et al., 2013).

The ability of a DNA sequence to bind a transcription factor is one of the simplest molecular phenotypes. It may evolve *de novo* or exaptively, from a binding site for a different transcription factor. For example, during the evolution of vertebrate αA-crystallin from a small heat shock protein, a transcription factor binding site known as a heat shock element was transformed into a binding site for the transcription factor Pax6, which helps drive crystalline expression in the eye lens (Cvekl et al., 2017). In the evolution of various cancers, mutations transform binding sites for the CCAAT-enhancer-binding protein (CEBP) into binding sites for multiple other transcription factors (Melton et al., 2015). Examples like these constitute the perhaps simplest possible molecular exaptations. Below I quantify from experimental DNA binding data how likely *de novo* or exaptive origins of new binding phenotypes are for 187 mouse transcription factors (Badis et al., 2009; Weirauch et al., 2014).

The second kind of novel phenotypes I study are metabolic phenotypes. More specifically, I study a metabolism's ability to sustain life – to synthesize all essential biomass molecules – in chemical environments that contain a novel source of carbon and energy. This ability comes about through changes in the metabolic genotype that specifies which enzyme-catalyzed reactions can take place in a metabolism. I take advantage of powerful and experimentally validated computational methods to predict metabolic phenotypes from genotypes (Edwards et al., 2001; Segre et al., 2002; Papp et al., 2004; Price et al., 2004; Feist et al., 2007). With these methods, I study whether metabolic phenotypes are more likely to originate *de novo* or exaptively. More generally, I show how information theory can help quantify the potential of a trait to originate *de novo* or exaptively. And I show that this potential strongly depends on the kind of trait considered.

Methods

Transcription factor binding data. I analyze a genotype space of 4^8 DNA molecules of length eight nucleotides and within this space those molecules bound by at least one of 187 mouse transcription factors. I use binding data from the UniPROBE (Newburger and Bulyk, 2009)(104 transcription factors) and the CIS-BP databases (Weirauch et al., 2014) (83 transcription factors) data bases. This data is based on high-throughput DNA binding experiments reported in (Badis et al., 2009) and (Weirauch et al., 2014), and was previously used in an analysis of DNA binding landscapes (Aguilar-Rodriguez et al., 2017). The data set includes a transcription factor if (i) its binding has been measured on two different kinds of protein binding micro arrays, and if (ii) it binds a sufficient number of sequences to permit an analysis of nucleotide interactions in binding. In addition, each factor to be included must bind at least one sequence with an E-score exceeding 0.45 (Aguilar-Rodriguez et al., 2017). The E-score is a proxy for a factor's relative binding affinity to a site. It is related to the Wilcoxon Mann Whitney test statistic, and ranges between -0.5 and +0.5 (strongest binding) (Badis et al., 2009). Because binding sites with E>0.35 are associated with a low false discovery rate of transcription factor binding (FDR=0.001), I here consider a site bound by any one factor if its E-score exceeds 0.35. I consider a non-palindromic binding site and its reverse complement as distinct sites.

Metabolic network analysis. My analysis begins with a small "universe" of 51 possible biochemical reactions from *E.coli* central carbon metabolism (Figure S1), and with ten different chemically minimal environments that differ only in the sole carbon and energy source they contain. While six of the 51 reactions are essential for the operation of central carbon metabolisms in all these environments, the presence of the remaining 45 reactions can vary without affecting viability in at least some environments (Hosseini et al., 2015). I focus on these 45 reactions, and thus analyze a metabolic genotype space whose 2^{45} members encode metabolic

reaction networks that are formed by all possible subsets of the 45 variable reactions (Hosseini et al., 2015). I define viability as the ability to synthesize those 13 biomass precursors (Figure S1) that are the starting points for the biosyntheses of some 60 biomass molecules that are essential to free-living microbes like *E.coli*, including 20 amino acids and four DNA nucleotide building blocks (Stryer, 1995; Noor et al., 2010). The carbon sources that distinguish the ten minimal environments are acetate, α-ketoglutarate, fumarate, fructose, glucose, glutamate, lactate, malate, pyruvate, and succinate(Hosseini et al., 2015). I note that a metabolism viable on any one carbon source may be viable on other carbon sources as well (Barve and Wagner, 2013; Hosseini and Wagner, 2016).

Concepts.

My approach focuses on the amount of genetic information that a genotype needs to harbor in order to specify or "encode" a phenotype, in the sense that the expression of this information produces the phenotype. All phenotypic traits are encoded by genotypes in some genotype space (Figure 1a). I will denote the size of this space by the number of genotypes $|G|$ in it. If every phenotype was specified by a single genotype, it would be trivial to quantify the amount of information needed to encode it. However, this is not the case. First, the same phenotype P (for example, an enzyme with a specific fold and catalytic activity) is typically specified by many genotypes. I denote the set of these genotypes by G_P. Second, the same genotype can also specify multiple phenotypes (Lipman and Wilbur, 1991; Schuster et al., 1994; Keefe and Szostak, 2001; Rodrigues and Wagner, 2009; Araya et al., 2012; Roscoe et al., 2013; Greenbury et al., 2014; Payne and Wagner, 2014).

The genotypes encoding any one phenotype P occupy some fraction $p=|G_P|/|G|$ of genotype space. The quantity $I_P := -log_2(p) = log_2|G| - log_2|G_P|$ can be viewed as the difference in Shannon entropy of two random variables that assume values $g \in G$ with probability $1/G$ and $g \in G_P$ with probabilities $1/G_P$, respectively (Cover and Thomas, 2006; Wagner, 2017). ($log_2(x)$ denotes the binary logarithm of x.) In various contexts, quantities analogous to I_P are called self-information, functional information, surprisal, and (biological) complexity (Adami et al., 2000; Carothers et al., 2004; Cover and Thomas, 2006; Wagner, 2017). I here refer to I_P as the informational complexity of the phenotype P (Adami et al., 2000; Carothers et al., 2004; Wagner, 2017). It ranges from a minimum of zero in the (extreme and trivial) case of $p = 1$, i.e., every genotype encodes the phenotype, to $log_2(|G|)$ if only one genotype in genotype space encodes the phenotype. The more genotypes encode a phenotype, the smaller is the informational complexity of this phenotype.

I will here focus on pairs of phenotypes, an old (ancestral) phenotype (P_O) and a new (derived) phenotype (P_N) that may emerge as an exaptation from the old phenotype or originate *de novo*. I will denote the fraction of genotype space that is occupied by genotypes encoding these two phenotypes as p_O and p_N, respectively. The amount of information that is needed to specify the new phenotype *de novo* is given by $I_N = -log_2(p_N)$. I want to compare this quantity to the amount of information needed if the new phenotype is to originate in an organism that already has the old phenotype. To compute the latter quantity, we can restrict ourselves to those genotypes that already encode the old phenotype P_O, which occupy a fraction p_O of genotype space. Among these genotypes, some fraction will also have the new phenotype, and I will denote this fraction as p_{ON} (Figure 1a). The amount of additional information required to encode phenotype P_N is equivalent to the proportion of genotypes with both old and new phenotypes, among all genotypes with the old phenotype. It is given by

$$I_{ex} = -log_2\left(\frac{p_{ON}}{p_O}\right). \quad (1)$$

and is related to a Kullback-Leibler distance, an elementary quantity from information theory (Cover and Thomas, 2006; Wagner, 2017).

To compare the amount of information needed for a *de novo* and an exaptive origin of P_N, we can compute the difference

$$\Delta I_{ex} \coloneqq I_N - I_{ex} = -log_2(p_N) - [-log_2\left(\frac{p_{ON}}{p_O}\right)] = log_2\left(\frac{p_{ON}}{p_O p_N}\right) \quad (2)$$

If this difference is positive, then more information is needed to specify the new phenotype *de novo* rather than starting from the old phenotype. In this sense, it is easier to evolve this phenotype from an existing phenotype. I note that I_N is the maximum value that ΔI_{ex} can assume, because $I_{ex} \geq 0$. That is, the maximal informational benefit that an old phenotype can provide for the origin of a new phenotype can be no greater than the amount of information needed to specify the new phenotype itself *de novo*. One can thus normalize ΔI_{ex}, dividing it by I_N to yield

$$\Delta I_{ex,n} = 1 - \frac{log_2\left(\frac{p_{ON}}{p_O}\right)}{log_2(p_N)} \quad (3)$$

This quantity will assume the maximal value of $\Delta I_{ex,n} = 1$ when it is informationally maximally advantageous to specify the new phenotype from the old phenotype, compared to specifying it *de novo*. To interpret the possible values of ΔI_{ex} and $\Delta I_{ex,n}$, a number of special cases need to be

distinguished that the following examples will illustrate. I will restrict myself to $p_{ON} > 0$ (Figure 1a), because it is the most relevant case for the exaptive origin of new phenotypes.

Results

Transcription factor binding sites. The experimental technology of protein binding microarrays can identify all short DNA sequences in a genotype space that a given transcription factor can bind (Badis et al., 2009; Weirauch et al., 2014). Briefly, this technology quantifies the binding of transcription factors to more than 10^4 different oligonucleotides of a given length immobilized on a microarray. The resulting data is already available for thousands of transcription factors (Badis et al., 2009; Newburger and Bulyk, 2009; Weirauch et al., 2014). I here analyze previously published data for the binding of 187 different mouse transcription factors to each such sequence in a genotype space of 4^8=65,536 DNA sequences (Badis et al., 2009; Weirauch et al., 2014; Aguilar-Rodriguez et al., 2017).

In the genotypes space of mouse transcription factors I study, individual transcription factors bind between 103 and 2933 sites, implying that between $-\log_2(2933/4^8) = 4.48$ and $-\log_2(103/4^8) = 9.31$ bits (depending on the transcription factor) need to be specified to encode a transcription factor binding phenotype (Wagner, 2017). Among all (187×186)/2=17391 possible pairs of transcription factors, these sets of binding sites overlap, i.e., $p_{ON} > 0$ for the vast majority (84.2%, 14645) of factor pairs. For these pairs, it is possible to transform one binding phenotype into another without transitioning through a site that is bound by a third factor or by no factor.

Figure 2a shows histograms of ΔI_{ex} and $\Delta I_{ex,n}$(inset). The distributions are slightly platykurtic, i.e., with fewer values around the mean, and fewer very large and small values than expected from a normal distribution (red line). The values of ΔI_{ex} range from -5.8 bits to +7.5 bits, with a slight excess of large values. Specifically, ΔI_{ex} lies between +5.8 and +7.5 bits ($\Delta I_{ex} > -\min\ \Delta I_{ex}$) for 70 TF pairs. Evolving a new binding specificity exaptively is especially advantageous for these TF pairs.

It is useful to distinguish three cases according to the sign of ΔI_{ex}, beginning with $\Delta I_{ex} > 0$. Consider the transformation of binding sites for SOX1, a transcription factor important in neurogenesis, into that for the related factor SRY (sex-determining region Y), which plays a role in gonad development (Guth and Wegner, 2008). For these DNA binding phenotypes, $\Delta I_{ex} = 4.62$ ($\Delta I_{ex,n} = 0.79$). This means that evolving a binding site for SRY from one for SOX1 requires 4.62 fewer bits – more than the information encoded in two base pairs – than evolving it *de novo*. The reason is that the set of binding sites for the two transcription overlap substantially.

Specifically, 42 percent (701) of the 1670 of binding sites of SOX1 are also binding sites for SRY.

SRY and SOX1 are from the same gene family (Guth and Wegner, 2008), and their high value of ΔI_{ex} is not a coincidence: Sets of binding sites overlap to the greatest extent for transcription factor pairs that originated through gene duplication (Weirauch et al., 2014). Some of the highest values of ΔI_{ex} come from such pairs. Among them are the genes encoding the odd-skipped-related transcription factors OSR1 and OSR2, which are involved in kidney, heart and palate development, and are duplicates with 65% amino acid sequence identity (Zhang et al., 2011). Evolving binding specificity for OSR2 from a binding site of OSR1 has ΔI_{ex}=6.45, i.e., it requires 6.45 fewer bits than evolving such specificity *de novo*. This value is not much lower than the information content of I_N=7.45 for an OSR1 binding site itself ($\Delta I_{ex,n} = 0.89$), indicating that most of the information encoding an OSR2-binding phenotype is already encoded in an OSR1-binding phenotype.

A second case is $\Delta I_{ex} < 0$, which means that it requires less information to evolve the new binding specificity *de novo* than from the old binding specificity. This will be the case if few DNA molecules can bind both transcription factors, such that p_{ON} is very small compared to p_O and p_N, which means that the amount of additional information I_{ex} to specify the new phenotype is very large.

One example involves the cell cycle regulators E2F3 and ARID3A (AT-rich interactive domain-containing protein 3A), members of a transcription factor family involved in hematopoiesis (Lees et al., 1993). $\Delta I_{ex} = -5.4$ bits for the evolution of E2F3 from ARID3A binding sites, whereas it takes only I_N=4.8 bits to specify a binding site for E2F3 *de novo*. It follows that $I_{ex} = 10.2$ bits, and that $\Delta I_{ex,n}$=-1.13, that is, it costs 113 percent more to evolve E2F3 binding from ARID3A binding than to evolve it *de novo*. The reason is that only two of the 2372 E2F3 binding sites are also binding sites for ARID3A ($I_{ex} = -\log(2/2372) = 10.21$). Starting from an ARID3A binding phenotype, much additional information is needed to evolve a binding site for E2F3.

More generally, as the fraction p_{ON} of genotypes that encode both phenotypes declines (and thus, as $-log_2\ p_{ON}$ increases), ΔI_{ex} also declines (Figure 2b). The two quantities are highly correlated (Spearman's r_s between $-log_2\ p_{ON}$ and ΔI_{ex}: r_s = - 0.85, p<10^{-17}, n=14645)

In a third scenario, $\Delta I_{ex} \approx 0$. That is, it is about equally likely that a binding site evolves *de novo* than that it evolves from the old binding sites. Equation (1) shows that this is the case if $p_O p_N \approx p_{ON}$. In other words, $\Delta I_{ex} \approx 0$ if the fraction p_{ON} of genotypes that bind both transcription

factors is equal to that expected by chance, i.e., if binding sites for the two transcription factors were independently distributed in genotype space. Examples where $\Delta I_{ex} \approx 0$ include ZIC1 (Zinc finger of the cerebellum) and SMAD3 (Mothers against decapentaplegic homolog 3), a regulator of cell proliferation (Zhang et al., 1996; Ali et al., 2012). Here, $\Delta I_{ex} = -0.0046$ bits, such that starting from a SMAD3 binding phenotype entails neither a strong advantage or disadvantage in evolving a binding site for ZIC1. Overall, 178 (0.01%) pairs of transcription factors have $|\Delta I_{ex}| < 0.05$, and figure 2a shows that the percentage of such transcription factors is smaller than expected if ΔI_{ex} followed a normal or binomial distribution.

Metabolic phenotypes. Metabolism is a complex network of enzyme-catalyzed chemical reactions that transforms environmental nutrients into small biomass molecules, such as amino acids and nucleotides. A metabolic genotype is that part of a genome's DNA that encodes all metabolic enzymes. Instead of representing such genotypes directly as DNA, metabolic systems analysis often represents them in a more abstract, reaction-centered way, as the complement of biochemical reactions that an organism's metabolism can catalyze (Edwards and Palsson, 2000; Schellenberger et al., 2010). This reaction-centered representation is especially appropriate wherever the elementary events of genetic change affect not just single nucleotides, but the presence or absence of enzyme-coding genes, such as in horizontal gene transfer and DNA recombination. Because the metabolic reactions encoded by any one genotype are a subset of a much larger "universe" of biochemical reactions, one can represent any organism's metabolic genotype as a binary vector whose entries indicates the presence ('1') or absence ('0') of specific metabolic reactions (Figure 3a). For tractability, I here analyze the metabolic genotype space of central metabolism, whose 2^{45} members encode metabolic networks that are formed by all possible subsets of 45 chemical reactions that can vary without necessarily abolishing an organism's viability in some environment (Hosseini et al., 2015). I note that the reaction complement of central carbon metabolisms also varies among organisms in the wild (Danson, 1989; Romano and Conway, 1996; Huynen et al., 1999). If $|G_P|$ denotes the total number of genotypes (subsets of chemical reactions) that convey viability in a given environment, then $-log_2(|G_P|/2^{45})$ quantifies the informational complexity of this viability phenotype. If a specific set of chemical reactions convey viability, then any superset of this set will do so as well (Hosseini et al., 2015).

Earlier work has used the experimentally validated computational method of flux balance analysis (FBA) to predict metabolic phenotypes for all 2^{45}central carbon genotypes (Hosseini et al., 2015). These phenotypes are viability phenotypes. That is, they reflect a metabolism's ability to produce biomass precursors essential for viability from a *single* source of carbon and energy, such as glucose (see Methods). Here I am considering ten such phenotypes (Figure 3b), namely

viability on acetate (ace), α-ketoglutarate (akg), fumarate (fum), fructose (fru), glucose (glc), glutamate (glu), lactate (lac), malate (mal), pyruvate (pyr), and succinate (suc). A new phenotype conveys viability on a new carbon source, and can come about through genetic changes, such as the addition of new enzyme-coding genes to a genome as a result of horizontal gene transfer and recombination (Hosseini et al., 2016).

Between I_{glc}=14.47 bits (glucose) and I_{ace}=21.6 bits (acetate), or, equivalently, metabolic reactions, are necessary to ensure viability on any one of the 10 carbon sources I consider. Figure 3c shows the distributions of ΔI_{ex} and $\Delta I_{ex,n}$ (inset). One noteworthy difference to the transcription factor binding phenotypes I considered above is that both quantities are always greater than zero, regardless of which pair of old and new phenotypes one considers. This means that it always requires more information to specify a new metabolic phenotype *de novo* than starting from an old phenotype. However, ΔI_{ex} still varies broadly, from 9.83 to 16.9 bits. The minimal value of ΔI_{ex} =9.83 bits (I_{ex}=4.64 bits) holds for the transition from metabolisms viable on α-ketoglutarate to metabolisms viable on glucose. It means that if a metabolism is already viable on α-ketoglutarate, one needs to specify on average 9.83 fewer bits to render it viable on glucose than when specifying a metabolism viable on glucose *de novo*. For this example, $\Delta I_{ex,n}$=0.68. That is, viability on α-ketoglutarate already provides 68% of the necessary information to specify viability on glucose. At the other extreme is the value of ΔI_{ex}=16.9 (I_{ex}=0.39) bits, which are necessary for a metabolism viable on succinate to originate from one viable on acetate instead of *de novo*. The corresponding value of $\Delta I_{ex,n}$=0.98 shows that it is close to maximally advantageous if the new phenotype originates exaptively rather than *de novo*.

When different new phenotypes are created from the same old phenotype, ΔI_{ex} can also vary among these phenotypes. For example, starting out from viability on fructose, the informational advantage of creating any one of the other nine phenotypes exaptively ranges between ΔI_{ex}=9.9 bits (α-ketoglutarate, $\Delta I_{ex,n}$=0.65) and 14.47 bits (glucose, $\Delta I_{ex,n}$=1). Similarly, the amount of additional information I_{ex} one needs to specify to obtain a new phenotype is low for some new phenotypes (malate from fructose: I_{ex}=3.1 bits) and much higher for others (acetate from fructose: I_{ex}=8.7 bits). Not surprisingly, the informationally expensive new phenotypes tend to be those that require many biochemical reactions, i.e., the phenotypes that are informationally complex (Spearman's r_s=0.42 between I_N and I_{ex}, and Spearman's r_s=0.53 between I_N and ΔI_{ex}; $p<3.6\times10^{-5}$; n=90).

Not just the new phenotype, but also the old, original phenotype can influence the informational advantage ΔI_{ex} of exaptation and the amount of information that needs to be specified for the new phenotype I_{ex}. Figure 3d shows the means (standard errors) of $\Delta I_{ex}$and I_{ex} (inset) for all

nine novel phenotypes starting from each of the 10 old phenotypes I consider. The informational advantage of exaptation clearly is greater for some starting phenotypes (e.g., acetate) than for others (glucose). The greater the informational complexity I_O of the starting phenotype, the smaller the number of additional bits (reactions) that need to be specified, and the greater the informational advantage of exaptation over *de novo* evolution of the new phenotype. (Spearman's r_s= -0.57 between I_O and I_{ex}, and Spearman's r_s=0.53 between I_O and ΔI_{ex}; $p<8\times10^{-8}$; n=90).

Discussion

My analysis of new phenotype origins is centered on transitions between phenotypes and not genotypes, because most phenotypes are encoded by multiple genotypes. Differences in the genetic information needed to encode a new and an old phenotype are well-suited to understand the likelihood of transitions between specific phenotypes, because they correlate with the amount of genetic change such transitions would require.

I considered two very different classes of phenotypes whose potential to originate exaptively differs. The first is the ability of DNA to bind a specific transcription factor, brought forth by specific sequences of nucleotides on a linear DNA string. The second is a metabolism's ability to synthesize essential molecules in a specific environment. It is brought forth by a chemical reaction network. A *de novo* origin is informationally cheaper than an exaptive origin for many transcription factor binding phenotypes. This does not hold for metabolic phenotypes, where an exaptive origin is always cheaper.

The cheaper exaptive origin of metabolic phenotypes has two reasons. The first is that metabolic phenotypes are more complex, even in the simple systems I study, requiring up to 21 bits compared to the maximally 9 bits of transcription factor binding. Second, and more importantly, the two traits have very different architectures. Metabolism can be partitioned into two parts. The first is a conserved core which comprises chemical reactions (enzymes) and pathways that are shared by many organisms, and that are needed for fundamental processes such as biosyntheses and energy conversion. The second is a more variable periphery, which comprises reactions and pathways responsible for viability in specific environments. Viability in new environments is primarily caused by alterations to this periphery (Pal et al., 2005; Bilgin and Wagner, 2012). What is more, as little as one or two reactions in the periphery usually suffice to bring forth viability in a new environment (Bilgin and Wagner, 2012). In other words, novel metabolic phenotypes result from the co-option of a metabolic core for a new role through the addition of novel metabolic reactions. This core-periphery structure has no counterpart in transcription factor binding.

In the transcription factor binding phenotypes I study, two bits correspond to one completely specified nucleotide. In the more abstract, reaction-based representation I use for metabolic phenotypes, one bit corresponds to one biochemical reaction. The two representations are not directly comparable. In a DNA-based representation of metabolic genotypes, every metabolic reaction would require a stretch of DNA that encodes the necessary enzyme, and thus requires much more than one bit of information. (Although the information content of some protein-based phenotypes, such ATP-binding, has been estimated (Keefe and Szostak, 2001; Carothers et al., 2004), this information content is unknown for most enzymes.) However when recombination or horizontal gene transfer are equally or more important than point mutations in creating novel phenotypes, a reaction-centered representation of information may be appropriate. This is indeed frequently the case, and especially in bacteria, where novel metabolic phenotypes often arise through horizontal transfer of enzyme-coding genes into a bacterial genome (Ochman et al., 2000; Pal et al., 2005; Copley, 2009). Importantly, a change in metabolic genotype representation would not affect the central conclusions of my analysis. For example, it would not change the fact that few reaction changes are needed to create new metabolic phenotypes. It would also not change the generally exaptive origin of novel metabolic phenotypes, nor would it change the observation that some phenotypes are informationally more expensive to create than others.

In the model of central carbon metabolism I consider, all central carbon metabolisms viable on fructose are also viable on glucose (Supplementary Results). Thus, viability on glucose is informationally 'free' – it requires no additional information ($I_{ex} = 0$), because it is an inevitable, latent by-product of viability on fructose. Not all latent traits are inevitable, because for many of the phenotypes I consider, only some – but not all – genotypes encode both an old (adaptive) and a new phenotype, that is, they lie in the intersection of the sets of genotypes encoding each trait (Figure 1a). In an organism with such a genotype the new phenotype may be latent. One can quantify the likelihood that an organism with an existing phenotype P_O also harbors a "new" and possibly latent phenotype P_N as p_{ON}/p_O, i.e., as the fraction of genotypes encoding *O* that also encode *N*.

Latent traits of both kinds – inevitable or not – are especially frequent in metabolism (Khersonsky and Tawfik, 2010; Nam et al., 2012; Barve and Wagner, 2013). They result both from the promiscuous activity of enzymes (Khersonsky and Tawfik, 2010; Nam et al., 2012), and from linear pathways of chemical reactions in parts of metabolism, where the ability to use one nutrient entails the ability to use others downstream of it (Barve and Wagner, 2013; Hosseini and Wagner, 2016). Latent phenotypes can also occur in the transcription factor binding phenotypes I study. Specifically, they are possible for those 84% of transcription factor pairs where $p_{ON} >$

0 (Figure 1a), and where regulatory cross-talk between two transcription factors can thus take place. For both metabolic and regulatory traits, the potential for a phenotype P_O to harbor a latent phenotype P_N varies widely.

The information theoretic framework I use creates concrete predictions. One of them is that phenotypes where $\Delta I_{ex}<0$ should more often originate *de novo* than by co-option from an existing phenotype P_O. This prediction is in principle testable, by analyzing the evolutionary origins of novel transcription factor binding sites through comparative genomics. A second prediction regards the observation that many novel traits – from aerobic respiration and nitrogen fixation to eye lenses and dissected leaves – have originated multiple times in evolution (Vermeij, 2006). All else being equal, I predict that phenotypes with low information content, or with little required additional information when they evolve exaptively, will have originated more often in life's evolution than informationally more complex phenotypes. A third prediction is that *de novo* origins should become progressively rarer as phenotypes become more complex, because it is informationally so much cheaper to evolve them as exaptations from existing phenotypes.

One limitation of my analysis is that I do not consider how easily a series of DNA mutations can access the set of genotypes encoding a new phenotype P_N from genotypes encoding and old phenotype P_O. Both sets of genotypes might be highly fragmented in genotype space, and only some of the subsets encoding phenotype P_O may intersect with a set of genotypes encoding P_N. If so, exaptive evolution of a new phenotype may be unlikely. For the transcription factor binding phenotypes and the metabolic phenotypes I consider here, this is not the case, because such fragmentation is very limited or nonexistent (Hosseini et al., 2015; Aguilar-Rodriguez et al., 2017). Fragmentation may exist for other traits, but a mix of computational analysis and experimental data shows for systems as different as evolving proteins and RNA molecules, regulatory circuits, and genome scale metabolisms, that genotypes encoding the same phenotype do not usually form highly fragmented sets but are instead connected in large networks that reach far through genotype space (Schuster et al., 1994; Schultes and Bartel, 2000; Ciliberti et al., 2007; Hayden et al., 2011; Greenbury et al., 2014; Payne and Wagner, 2014). Where fragmentation exists, it can sometimes be overcome through a combination of genetic drift, high mutation rates, or large population sizes (Weinreich and Chao, 2005; Weissman et al., 2009).

A second limitation is the tacit assumption that all genotypes encoding a phenotype are equivalent, but in practice they are not. For example, the metabolic enzymes encoded by different variants of the same gene may differ in their catalytic rate or thermodynamic stability, and these differences can lead to fitness differences. Unfortunately, we currently lack

quantitative fitness data on large enough numbers of genotypes to incorporate such data into any of the examples I discuss. However, when such data become available, the theoretical framework I propose will be able to accommodate it. For example, one can weight different genotypes in calculations of informational complexity according to their likelihood of occurrence in a population, which may depend on their fitness.

A third limitation is that I have focused on systems where genotype space is sufficiently small to be enumerated exhaustively. For informationally more complex phenotypes this is not going to be possible. It is tempting to overcome this limitation through the simplifying assumption that individual system parts – here, nucleotides or biochemical reactions – contribute additively to a phenotype (Adami et al., 2000). However, doing so can lead to substantial underestimates of phenotypic information content (Wagner, 2017). An alternative is to sample sequence space in ways suitable to infer phenotypic complexity. However, statistical methods to do so are still in their infancy (Wagner, 2017). Our dramatically increasing ability to genotype and phenotype large populations of organisms will hopefully prompt the development of such methods. Whether they will ever allow us to apply information theory to phenotypes as complex as our middle ear bones is an open question. But information theory can undoubtedly help us turn the analysis of trait origins into a quantitative science suitable to study phenotypes of ever-increasing complexity.

Acknowledgments

I would like to acknowledge support from the European Research Council under Grant Agreement No. 739874, as well as from Swiss National Science Foundation grant 31003A_172887, and the University Priority Research Program in Evolutionary Biology at the University of Zurich**.**

Figure Captions

Figure 1. Various possible relationships among sets of genotypes with the same phenotype. Large squares symbolize genotype space, circles correspond to sets of genotypes with the same phenotype, which can be either an old phenotype P_O (set *O*), or a new phenotype P_N (set *N*). **a)** The two genotype sets have a non-empty intersection (denoted by $O \cap N$), which is the main scenario I consider. **b)** The two genotype sets have an empty intersection (p_{ON}=0), which makes exaptation unlikely or impossible. **c)** The set of genotype with the old phenotype P_O is a subset of that with the new phenotype P_N. In this case, the new phenotype is an inevitable, possibly latent by-product of the old phenotype.

Figure 2: The informational cost of exaptation in transcription factor binding phenotypes. **a)** Histogram of ΔI_{ex} (in bits) and $\Delta I_{ex,n}$ (inset) for 14645 pairs of transcription factors where p_{ON}>0. The red line shows a Gaussian distribution. **b)** ΔI_{ex} (in bits) as a function of the negative binary logarithm of p_{ON}, the fraction of genotypes (transcription factor binding sites) that can bind both factors in a pair. The black line is a linear regression line.

Figure 3. Metabolic exaptations. a) The genotype encoding a network of metabolic reactions can be represented through the presence (black type, '1') or absence (grey type, 0') of individual metabolic reactions, which are represented in this hypothetical example through their stoichiometric equations. In this analysis I consider 45 reactions from central carbon metabolism whose presence can vary and still allow viability. **b)** The environments I consider here are chemically minimal environments (Hosseini et al., 2015) that differ only in one of the 10 carbon sources shown here. A carbon source is shown as supporting viability (black type, '1') if a metabolism can synthesize each of 12 different biomass precursors (Figure S1) when this carbon source is the only carbon source. **c)** Distribution of ΔI_{ex} (in bits) and the normalized $\Delta I_{ex,n}$ (inset) for the 90 pairs of carbon sources that can be formed by the 10 carbon sources I consider here. **d)** Mean (circles) and standard error of the mean (whiskers) for ΔI_{ex} (in bits) and I_{ex} (inset) for the nine novel phenotypes that can be formed from each of the 10 old phenotypes shown on the horizontal axis.

Literature Cited

Adami, C., Ofria, C., and Collirer, T.C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences* 97**,** 4463-4468.

Aguilar-Rodriguez, J., Payne, J.L., and Wagner, A. (2017). 1000 empirical adaptive landscapes and their navigability. *Nature Ecology and Evolution* 1**,** 0045.

Ali, R.G., Bellchambers, H.M., and Arkell, R.M. (2012). Zinc fingers of the cerebellum (Zic): transcription factors and co-factors. *The international journal of biochemistry & cell biology* 44(11)**,** 2065-2068.

Araya, C.L., Fowler, D.M., Chen, W.T., Muniez, I., Kelly, J.W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America* 109(42)**,** 16858-16863. doi: 10.1073/pnas.1209751109.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935)**,** 1720-1723. doi: 10.1126/science.1162327.

Barve, A., and Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* 500**,** 203-206.

Bilgin, T., and Wagner, A. (2012). Design constraints on a synthetic metabolism. *PLoS One* 7(6). doi: e39903 10.1371/journal.pone.0039903.

Blount, Z.D., Borland, C.Z., and Lenski, R.E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* 105(23)**,** 7899-7906. doi: 10.1073/pnas.0803151105.

Carothers, J.M., Oestreich, S.C., Davis, J.H., and Szostak, J.W. (2004). Informational complexity and functional activity of RNA structures. *Journal of the American Chemical Society* 126(16)**,** 5130-5137. doi: 10.1021/ja031504a.

Ciliberti, S., Martin, O.C., and Wagner, A. (2007). Circuit topology and the evolution of robustness in complex regulatory gene networks. *PLoS Computational Biology* 3(2): e15.

Copley, S.D. (2009). Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nature Chemical Biology* 5(8)**,** 560-567. doi: 10.1038/nchembio.197.

Cover, T.M., and Thomas, J.A. (2006). *Elements of information theory.* Wiley: Hoboken, New Jersey.

Cvekl, A., Zhao, Y.L., McGreal, R., Xie, Q., Gu, X., and Zheng, D.Y. (2017). Evolutionary origins of Pax6 control of crystallin genes. *Genome Biology and Evolution* 9(8)**,** 2075-2092. doi: 10.1093/gbe/evx153.

Danson, M.J. (1989). Central metabolism of the archaebacteria: an overview. *Canadian Journal of Microbiology* 35(1)**,** 58-64.

Darwin, C. (1872). *The origin of species by means of natural selection, or the preservation of favored races in the struggle for life (6th ed., reprinted by A.L. Burt, New York).* London, England: John Murray

de Visser, J.A.G.M., and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* 15(7)**,** 480-490. doi: 10.1038/nrg3744.

Dhar, R., Sägesser, R., Weikert, C., and Wagner, A. (2013). Yeast adapts to a changing stressful environment by evolving cross-protection and anticipatory gene regulation. *Molecular Biology and Evolution* 30(3)**,** 573-588.

Dhar, R., Sagesser, R., Weikert, C., Yuan, J., and Wagner, A. (2011). Adaptation of Saccharomyces cerevisiae to saline stress through laboratory evolution. *Journal of Evolutionary Biology* 24(5)**,** 1135-1153. doi: 10.1111/j.1420-9101.2011.02249.x.

Edwards, J.S., Ibarra, R.U., and Palsson, B.O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nature Biotechnology* 19(2)**,** 125-130.

555 Edwards, J.S., and Palsson, B.O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: Its
556 definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of*
557 *the United States of America* 97(10)**,** 5528-5533.
558 Fay, R.R., Manley, G.A., and Popper, A.N. (2004). *Evolution of the vertebrate auditory system.* Berlin:
559 Springer.
560 Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., et al. (2007). A
561 genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260
562 ORFs and thermodynamic information. *Molecular Systems Biology* 3. doi:
563 12110.1038/msb4100155.
564 Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., and Carroll, S.B. (2005). Chance caught on
565 the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*
566 433(7025)**,** 481-487. doi: 10.1038/nature03235.
567 Gould, S., and Vrba, E. (1982). Exaptation - a missing term in the science of form. *Paleobiology* 8(1)**,** 4-
568 15.
569 Greenbury, S.F., Johnston, I.G., Louis, A.A., and Ahnert, S.E. (2014). A tractable genotype–phenotype
570 map modelling the self-assembly of protein quaternary structure. *Journal of the Royal Society*
571 *Interface* 11(95)**,** 20140249.
572 Guerreiro, I., Nunes, A., Woltering, J.M., Casaca, A., Novoa, A., Vinagre, T., et al. (2013). Role of a
573 polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine.
574 *Proceedings of the National Academy of Sciences of the United States of America* 110(26)**,**
575 10682-10686. doi: 10.1073/pnas.1300592110.
576 Guth, S., and Wegner, M. (2008). Having it both ways: Sox protein function between conservation and
577 innovation. *Cellular and molecular life sciences* 65(19)**,** 3000-3018.
578 Hayden, E.J., Ferrada, E., and Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary
579 adaptation in an RNA enzyme. *Nature* 474(7349)**,** 92-95. doi: 10.1038/nature10083.
580 Hosseini, S.-R., Barve, A., and Wagner, A. (2015). Exhaustive analysis of a genotype space comprising
581 10^{15} central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLoS*
582 *Computational Biology* 11(8)**,** e1004329.
583 Hosseini, S.-R., Martin, O.C., and Wagner, A. (2016). Phenotypic innovation through recombination in
584 genome-scale metabolic networks. *Proceedings of the Royal Society Series B* 283**,** 20161536.
585 Hosseini, S.-R., and Wagner, A. (2016). The potential for non-adaptive origins of evolutionary
586 innovations in central carbon metabolism. *BMC Systems Biology* 10(1)**,** 97.
587 Huynen, M.A., Dandekar, T., and Bork, P. (1999). Variation and evolution of the citric acid cycle: a
588 genomic perspective. *Trends in Microbiology* 7**,** 281-291.
589 Keefe, A.D., and Szostak, J.W. (2001). Functional proteins from a random-sequence library. *Nature*
590 410(6829)**,** 715-718.
591 Khersonsky, O., and Tawfik, D.S. (2010). Enzyme promiscuity: A mechanistic and evolutionary
592 perspective. *Annual Review of Biochemistry* 79**,** 471-505. doi: 10.1146/annurev-biochem-030409-
593 143718.
594 Lees, J.A., Saito, M., Vidal, M., Valentine, M., Look, T., Harlow, E., et al. (1993). The retinoblastoma
595 protein binds to a family of E2F transcription factors. *Molecular and cellular biology* 13(12)**,**
596 7813-7825.
597 Lipman, D., and Wilbur, W. (1991). Modeling neutral and selective evolution of protein folding.
598 *Proceedings of the Royal Society of London Series B* 245(1312)**,** 7-11.
599 Luo, Z.X. (2007). Transformation and diversification in early mammal evolution. *Nature* 450(7172)**,**
600 1011-1019. doi: 10.1038/nature06277.
601 Melton, C., Reuter, J.A., Spacek, D.V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory
602 regions of human cancer genomes. *Nature Genetics* 47(7)**,** 710-716. doi: 10.1038/ng.3332.
603 Nam, H., Lewis, N.E., Lerman, J.A., Lee, D.H., Chang, R.L., Kim, D., et al. (2012). Network context and
604 selection in the evolution to enzyme specificity. *Science* 337(6098)**,** 1101-1104. doi:
605 10.1126/science.1216861.

606 Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray
607 data on protein–DNA interactions. *Nucleic Acids Research* 37(Suppl 1)**,** D77-D82.
608 Noor, E., Eden, E., Milo, R., and Alon, U. (2010). Central carbon metabolism as a minimal biochemical
609 walk between precursors for biomass and energy. *Molecular Cell* 39(5)**,** 809-820.
610 Norell, M.A., and Xu, X. (2005). Feathered dinosaurs. *Annual Review of Earth and Planetary Sciences*
611 33**,** 277-299. doi: 10.1146/annurev.earth.33.092203.122511.
612 Ochman, H., Lawrence, J., and Groisman, E. (2000). Lateral gene transfer and the nature of bacterial
613 innovation *Nature* 405**,** 299-304.
614 Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic networks by
615 horizontal gene transfer. *Nature Genetics* 37(12)**,** 1372-1375. doi: 10.1038/ng1686.
616 Palmer, A.C., Toprak, E., Baym, M., Kim, S., Veres, A., Bershtein, S., et al. (2015). Delayed commitment
617 to evolutionary fate in antibiotic resistance fitness landscapes. *Nature communications* 6. doi:
618 10.1038/ncomms8385.
619 Papp, B., Pal, C., and Hurst, L.D. (2004). Metabolic network analysis of the causes and evolution of
620 enzyme dispensability in yeast. *Nature* 429(6992)**,** 661-664.
621 Payne, J.L., and Wagner, A. (2014). The robustness and evolvability of transcription factor binding sites.
622 *Science* 343(6173)**,** 875-877.
623 Piatigorsky, J. (1998). Gene sharing in lens and cornea: Facts and implications. *Progress in Retinal and*
624 *Eye Research* 17(2)**,** 145-174.
625 Piatigorsky, J., and Wistow, G.J. (1989). Enzyme crystallins : Gene sharing as an evolutionary strategy.
626 *Cell* 57(#2)**,** 197-199.
627 Price, N., Reed, J., and Palsson, B. (2004). Genome-scale models of microbial cells: evaluating the
628 consequences of constraints. *Nature Reviews Microbiology* 2**,** 886-897.
629 Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory evolution.
630 *Proceedings of the National Academy of Sciences of the United States of America* 104**,** 8605-
631 8612. doi: 10.1073/pnas.0700488104.
632 Qasba, P.K., and Kumar, S. (1997). Molecular divergence of lysozymes and alpha-lactalbumin. *Critical*
633 *Reviews in Biochemistry and Molecular Biology* 32(4)**,** 255-306. doi:
634 10.3109/10409239709082574.
635 Rodrigues, J.F.M., and Wagner, A. (2009). Evolutionary plasticity and innovations in complex metabolic
636 reaction networks. *PLoS Computational Biology* 5(12). doi: 10.1371/journal.pcbi.1000613.
637 Romano, A., and Conway, T. (1996). Evolution of carbohydrate metabolic pathways. *Research in*
638 *Microbiology* 147(6-7)**,** 448-455.
639 Roscoe, B.P., Thayer, K.M., Zeldovich, K.B., Fushman, D., and Bolon, D.N.A. (2013). Analyses of the
640 effects of all ubiquitin point mutants on yeast growth rate. *Journal of Molecular Biology* 425(8)**,**
641 1363-1377. doi: 10.1016/j.jmb.2013.01.032.
642 Schellenberger, J., Park, J.O., Conrad, T.M., and Palsson, B.O. (2010). BiGG: a Biochemical Genetic and
643 Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 11. doi:
644 21310.1186/1471-2105-11-213.
645 Schultes, E., and Bartel, D. (2000). One sequence, two ribozymes: Implications for the emergence of new
646 ribozyme folds. *Science* 289(5478)**,** 448-452.
647 Schuster, P., Fontana, W., Stadler, P., and Hofacker, I. (1994). From sequences to shapes and back - a
648 case-study in RNA secondary structures. *Proceedings of the Royal Society of London Series B*
649 255(1344)**,** 279-284.
650 Segre, D., Vitkup, D., and Church, G. (2002). Analysis of optimality in natural and perturbed metabolic
651 networks. *Proceedings of the National Academy of Sciences of the U.S.A.* 99**,** 15112-15117.
652 Stryer, L. (1995). *Biochemistry.* New York: Freeman.
653 Thorington, R.W., and Santana, E.M. (2007). How to make a flying squirrel: Glaucomys anatomy in
654 phylogenetic perspective. *Journal of Mammalogy* 88(4)**,** 882-896. doi: 10.1644/06-mamm-s-
655 325r2.1.

Toll-Riera, M., San Millan, A., Wagner, A., and MacLean, R.C. (2016). The genomic basis of evolutionary innovation in Pseudomonas aeruginosa. *PLoS Genetics* 12(5)**,** e1006005.

True, J.R., and Carroll, S.B. (2002). Gene co-option in physiological and morphological evolution. *Annual Review of Cell and Developmental Biology* 18**,** 53-80. doi: 10.1146/annurev.cellbio.18.020402.140619.

Vermeij, G.J. (2006). Historical contingency and the purported uniqueness of evolutionary innovations. *Proceedings of the National Academy of Sciences of the United States of America* 103(6)**,** 1804-1809.

Wagner, A. (2017). Information theory, evolutionary innovations and evolvability. *Phil. Trans. R. Soc. B* 372(1735)**,** 20160416.

Weinreich, D.M., and Chao, L. (2005). Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* 59(6)**,** 1175-1182.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6)**,** 1431-1443.

Weissman, D.B., Desai, M.M., Fisher, D.S., and Feldman, M.W. (2009). The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology* 75(4)**,** 286-300. doi: 10.1016/j.tpb.2009.02.006.

Zhang, Y., Feng, X.-H., Wu, R.-Y., and Derynck, R. (1996). Receptor-associated Mad homologues synergize as effectors of the TGF-β response. *Nature* 383(6596)**,** 168.

Zhang, Z., Iglesias, D., Eliopoulos, N., El Kares, R., Chu, L., Romagnani, P., et al. (2011). A variant OSR1 allele which disturbs OSR1 mRNA expression in renal progenitor cells is associated with reduction of newborn kidney size and function. *Human Molecular Genetics* 20(21)**,** 4167-4174.

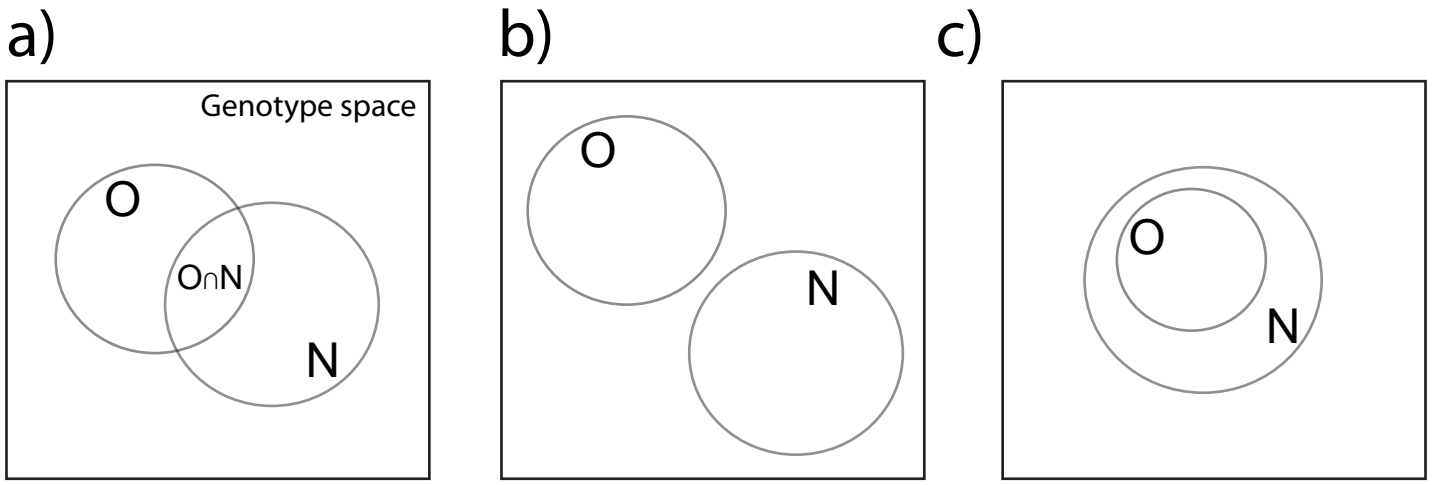

Figure 1

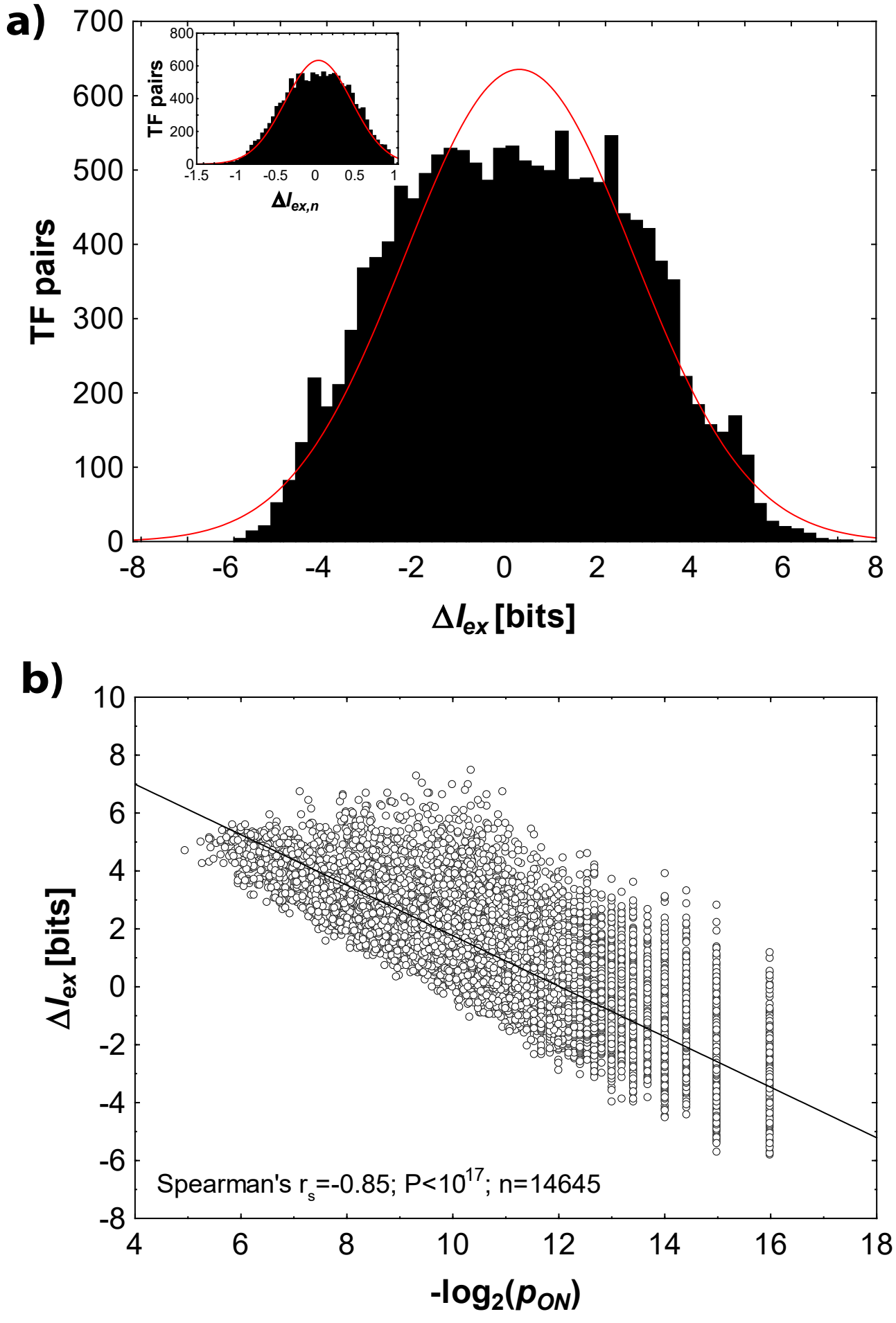
a)
TF pairs
ΔI_{ex} [bits]
$\Delta I_{ex,n}$
b)
ΔI_{ex} [bits]
$-\log_2(p_{ON})$
Spearman's r_s=-0.85; P<10^{17}; n=14645

Figure 2

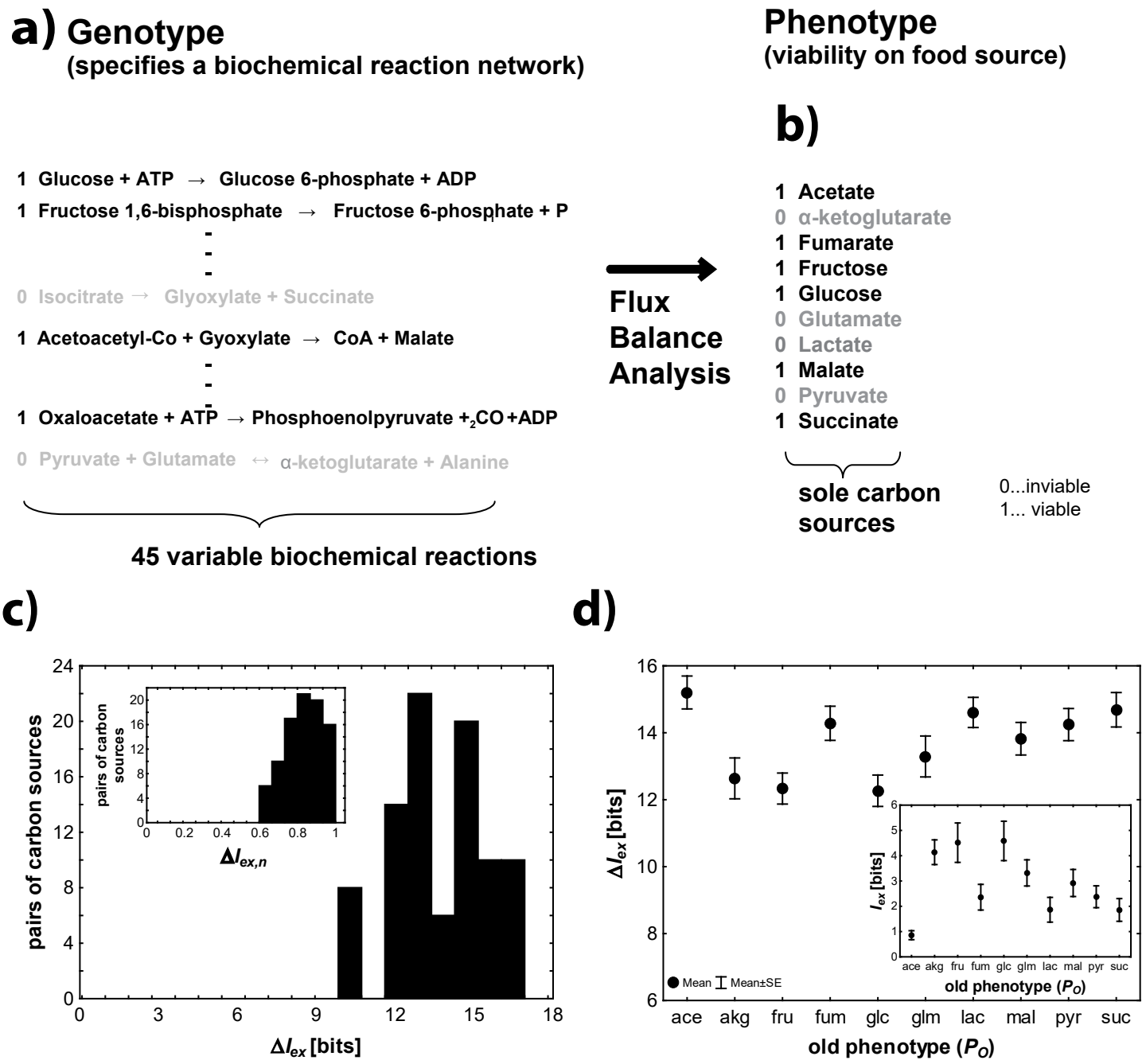
a) Genotype
(specifies a biochemical reaction network)
1 Glucose + ATP → Glucose 6-phosphate + ADP
1 Fructose 1,6-bisphosphate → Fructose 6-phosphate + P
0 Isocitrate → Glyoxylate + Succinate
1 Acetoacetyl-Co + Gyoxylate → CoA + Malate
1 Oxaloacetate + ATP → Phosphoenolpyruvate + CO2 + ADP
0 Pyruvate + Glutamate ↔ α-ketoglutarate + Alanine
45 variable biochemical reactions
Flux Balance Analysis
Phenotype
(viability on food source)
b)
1 Acetate
0 α-ketoglutarate
1 Fumarate
1 Fructose
1 Glucose
0 Glutamate
0 Lactate
1 Malate
0 Pyruvate
1 Succinate
sole carbon sources
0...inviable
1... viable
c)
pairs of carbon sources
ΔIex [bits]
ΔIex,n
d)
ΔIex [bits]
Iex [bits]
Mean
Mean±SE
ace akg fru fum glc glm lac mal pyr suc
old phenotype (PO)

Figure 3

Supplementary Information to

Information theory can help quantify the potential of new phenotypes to originate as exaptations

Andreas Wagner[a,b,c]

[a]University of Zurich, Institute of Evolutionary Biology and Environmental Studies, Zurich, Switzerland,

[b]Swiss Institute of Bioinformatics, Lausanne, Switzerland,

[c]Santa Fe Institute, Santa Fe, New Mexico, USA

Supplementary results

Latent phenotypes are inevitable by-products of other phenotypes that do not require additional genetic information to be specified. There are two pairs of carbon sources where $\Delta I_{ex,n}$ assumes its maximally possible value of one. These are fructose-glucose, as well as fumarate-malate. For the first pair, it is noteworthy that the amount of information needed to specify each phenotype is very similar (I_{fru}=14.54 bits, and I_{glc}=14.47 bits). More importantly, closer inspection reveals that all 1.47×10^9 genotypes that are viable on fructose are also viable on glucose (1.55×10^9 genotypes, Figure 1c). In other words, if a metabolism is viable on fructose already, it takes no additional information to turn it into a metabolism viable on glucose. Viability on glucose is an inevitable, latent by-product of viability on fructose, because glucose and fructose are biochemically similar and metabolized by similar pathways. It is not surprising then that ΔI_{ex} =14.47 bits, exactly the same amount of information needed to specify viability on glucose *de novo.* This explains the maximal possible value of $\Delta I_{ex,n}$=1. I note that the converse does not hold, that is, turning a metabolism that is viable on glucose into one that is viable on fructose requires additional information, because not all metabolisms viable on glucose are also viable on fructose. However, this additional information is quite small (I_{ex}=0.07 bits), because the vast majority of metabolisms viable on glucose (94.8%) are already viable on fructose. More specifically, ΔI_{ex}=14.47 bits, which is slightly lower than the possible maximum of I_N=14.54 for fructose, yielding $\Delta I_{ex,n}$=0.995. The same qualitative patterns exist for fumarate and malate, i.e., all metabolisms viable on fumarate are also viable on malate. In sum, even simple metabolic networks can harbor latent phenotypic traits whose origin does not require any additional genetic information.

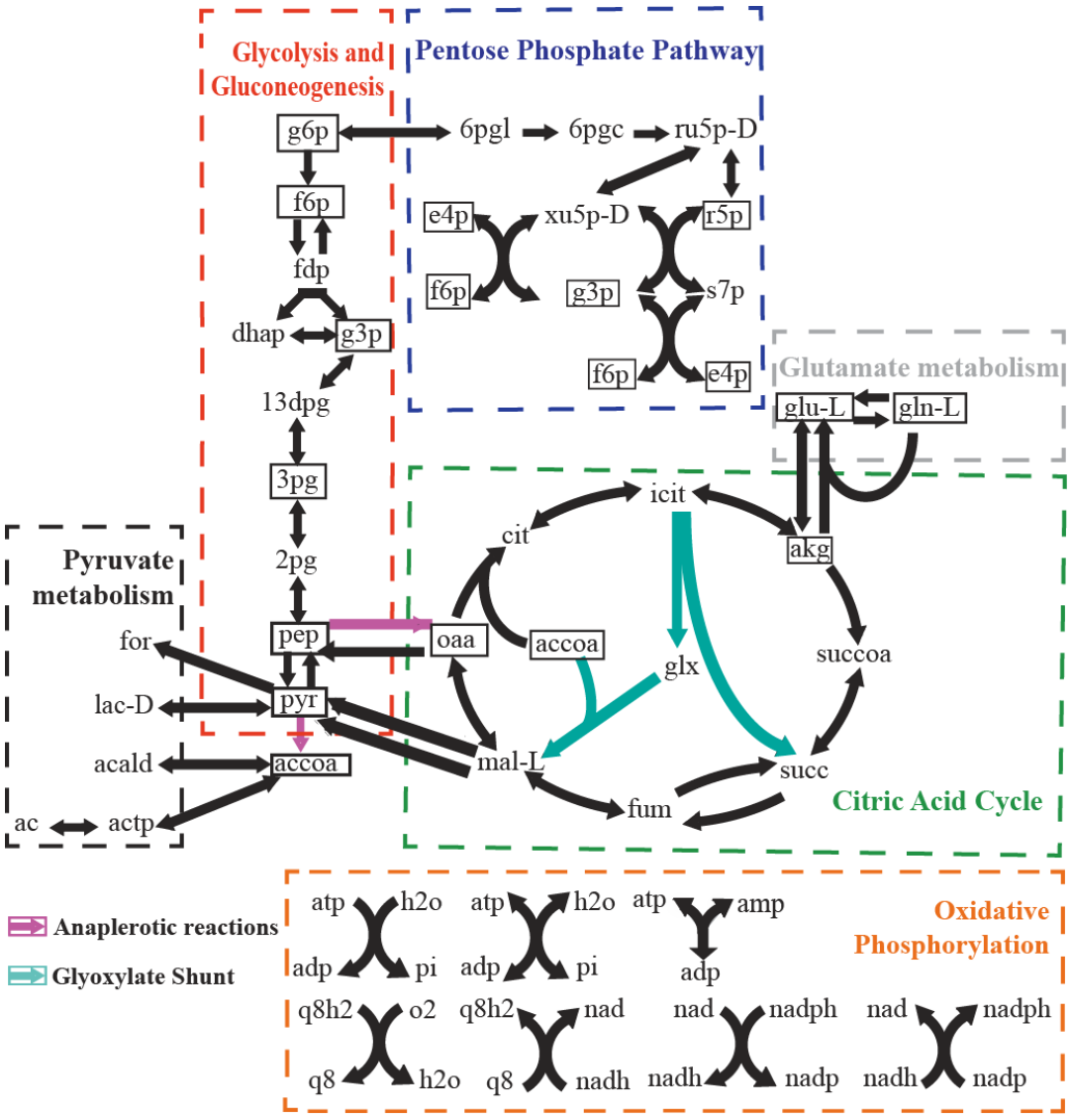

Figure S1: Central carbon metabolism. Figure and captions with modification from (Hosseini et al., 2015). Each arrow corresponds to one of the internal reactions considered here and in (Hosseini et al., 2015). Metabolites are indicated by acronyms (Hosseini et al., 2015). Boxed metabolites correspond to the 13 essential biomass precursors D-glucose-6-phosphate (g6p), D-fructose 6-phosphate (f6p), glyceraldehyde 3-phosphate (g3p), 3-phospho-D-glycerate (3pg), phosphoenolpyruvate (pep), pyruvate (pyr), acetyl CoA (accoa), D-eryhtrose 4-phosphate (e4p), D-ribose 5-phosphate (r5p), L-glutamine (gln-L), L-glutamate (glu-L), oxaloacetate (oaa), and 2-oxoglutarate (α-ketoglutarate, akg). Note that four of the precursors (accoa, g3p, f6p and e4p) are shown more than once for visual clarity. Six of the 51 arrows (reactions) shown here cannot be eliminated without abolishing viability in each of the environments considered here (Hosseini et al., 2015). Because such reactions are always essential I excluded them from all calculations.